
Minimum Probability Flow Learning

Jascha Sohl-Dickstein^{ad1*}, Peter Battaglino^{bd2*} and Michael R. DeWeese^{bed3}

^aBiophysics Graduate Group, ^bDepartment of Physics, ^cHelen Wills Neuroscience Institute

^dRedwood Center for Theoretical Neuroscience

University of California, Berkeley, 94720

¹jascha@berkeley.edu, ²pbb@berkeley.edu,

³deweese@berkeley.edu, ^{*}*These authors contributed equally.*

Abstract

Fitting probabilistic models to data is often difficult, due to the general intractability of the partition function and its derivatives. Here we propose a new parameter estimation technique that does not require computing an intractable normalization factor or sampling from the equilibrium distribution of the model. This is achieved by establishing dynamics that would transform the observed data distribution into the model distribution, and then setting as the objective the minimization of the KL divergence between the data distribution and the distribution produced by running the dynamics for an infinitesimal time. Score matching, minimum velocity learning, and certain forms of contrastive divergence are shown to be special cases of this learning technique. We demonstrate parameter estimation in Ising models, deep belief networks and a product of Student-t test model of natural scenes. In the Ising model case, current state of the art techniques are outperformed by approximately two orders of magnitude in learning time, with comparable error in recovered parameters. This technique promises to broaden the class of probabilistic models that are practical for use with large, complex data sets.

1 Introduction

Estimating parameters for probabilistic models is a fundamental problem in many scientific and engineering disciplines. Unfortunately, most probabilistic learning techniques require calculating the normalization factor, or partition function, of the probabilistic model in question, or at least calculating its gradient. For the overwhelming majority of models there are no known analytic solutions, confining us to the restrictive subset of probabilistic models that can be solved analytically, or those that can be made tractable using approximate learning techniques. Thus, development of powerful new techniques for parameter estimation promises to greatly expand the variety of models that can be fit to complex data sets.

Many approaches exist for approximate learning, including mean field theory and its expansions, variational Bayes techniques and a plethora of sampling or numerical integration based methods [23, 10, 9, 5]. Of particular interest are contrastive divergence (CD), developed by Hinton, Welling and Carreira-Perpiñán [24, 4], Hyvärinen’s score matching (SM) [7] and the minimum velocity learning framework proposed by Movellan [15, 14, 16].

Contrastive divergence [24, 4] is a variation on steepest gradient descent of the maximum (log) likelihood (ML) objective function. Rather than integrating over the full model distribution, CD approximates the partition function term in the gradient by averaging over the distribution obtained

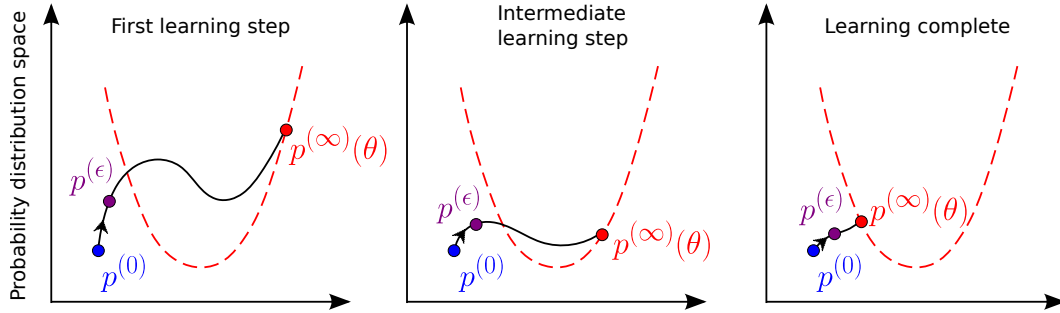


Figure 1: Dynamics are established which transform any initial distribution into the model distribution $\mathbf{p}^{(\infty)}(\theta)$. The dashed red line indicates the family of distributions parametrized by θ , and the three successive figures illustrate graphically the progression of learning. Under maximum likelihood learning, model parameters θ are chosen so as to minimize the Kullback–Leibler divergence between the data distribution $\mathbf{p}^{(0)}$ and the model distribution $\mathbf{p}^{(\infty)}(\theta)$. Under minimum probability flow learning, however, the KL divergence between $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(\epsilon)}$ is minimized instead, where $\mathbf{p}^{(\epsilon)}$ is the distribution obtained by evolving $\mathbf{p}^{(0)}$ for infinitesimal time under the dynamics. Here we represent graphically how pulling $\mathbf{p}^{(\epsilon)}$ as close as possible to $\mathbf{p}^{(0)}$ tends to pull $\mathbf{p}^{(\infty)}(\theta)$ close to $\mathbf{p}^{(0)}$ as well.

after taking a few Markov chain Monte Carlo (MCMC) steps away from the data distribution¹. Qualitatively, one can imagine that the data distribution is contrasted against a distribution which has evolved only a small distance towards the model distribution, whereas it would be contrasted against the true model distribution in traditional MCMC approaches. Although CD is not guaranteed to converge to the right answer, or even to a fixed point, it has proven to be an effective and fast heuristic for parameter estimation [12, 26].

Score matching, developed by Aapo Hyvärinen [7], is a method that learns parameters in a probabilistic model using only derivatives of the energy function evaluated over the data distribution (see Equation (17)). This sidesteps the need to explicitly sample or integrate over the model distribution. In score matching one minimizes the expected square distance of the score function with respect to spatial coordinates given by the data distribution from the similar score function given by the model distribution. A number of connections have been made between score matching and other learning techniques [8, 22, 15, 11].

Minimum velocity learning is an approach recently proposed by Movellan [15] that recasts a number of the ideas behind CD, treating the minimization of the initial dynamics away from the data distribution as the goal itself rather than a surrogate for it. Movellan’s proposal is that rather than directly minimize the difference between the data and the model, one introduces system dynamics that have the model as their equilibrium distribution, and minimizes the initial flow of probability away from the data under those dynamics. If the model looks exactly like the data there will be no flow of probability, and if model and data are similar the flow of probability will tend to be minimal. Movellan applies this intuition to the specific case of distributions over continuous state spaces evolving via diffusion dynamics, and recovers the score matching objective function. The velocity

¹ The update rule for gradient descent of the negative log likelihood, or maximum likelihood objective function, is

$$\Delta\theta \propto \frac{\partial \left[\sum_i p_i^{(0)} \log p_i^{(\infty)}(\theta) \right]}{\partial \theta} = - \sum_i \frac{\partial E_i(\theta)}{\partial \theta} p_i^{(0)} + \sum_i \frac{\partial E_i(\theta)}{\partial \theta} p_i^{(\infty)}(\theta),$$

where $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(\infty)}(\theta)$ represent the data distribution and model distribution, respectively, $\mathbf{E}(\theta)$ is the energy function associated with the model distribution and i indexes the states of the system (see Section 2.1). The second term in this gradient can be extremely difficult to compute (costing in general an amount of time exponential in the dimensionality of the system). Under contrastive divergence $p_i^{(\infty)}(\theta)$ is replaced by samples only a few Monte Carlo steps away from the data.

in minimum velocity learning is the difference in average drift velocities between particles diffusing under the model distribution and particles diffusing under the data distribution.

Here we propose a framework called minimum probability flow learning (MPF), applicable to *any* parametric model, of which minimum velocity, SM and certain forms of CD are all special cases, and which is in many situations more powerful than any of these algorithms. We demonstrate that learning under this framework is effective and fast in a number of cases: Ising models [3, 1], deep belief networks [6], and the product of Student-t tests model for natural scenes [25].

2 Minimum probability flow

Our goal is to find the parameters that cause a probabilistic model to best agree with a set \mathcal{D} of (assumed iid) observations of the state of a system. We will do this by proposing dynamics that guarantee the transformation of the data distribution into the model distribution, and then minimizing the KL divergence which results from running those dynamics for a short time ϵ (see Figure 1).

2.1 Distributions

The data distribution is represented by a vector $\mathbf{p}^{(0)}$, with $p_i^{(0)}$ the fraction of the observations \mathcal{D} in state i . The superscript (0) represents time $t = 0$ under the system dynamics (which will be described in more detail in Section 2.2). For example, in a two variable binary system, $\mathbf{p}^{(0)}$ would have four entries representing the fraction of the data in states 00, 01, 10 and 11 (Figure 2).

Our goal is to find the parameters θ that cause a model distribution $\mathbf{p}^{(\infty)}(\theta)$ to best match the data distribution $\mathbf{p}^{(0)}$. The superscript (∞) on the model distribution indicates that this is the equilibrium distribution reached after running the dynamics for infinite time. Without loss of generality, we assume the model distribution is of the form

$$p_i^{(\infty)}(\theta) = \frac{\exp(-E_i(\theta))}{Z(\theta)}, \quad (1)$$

where $\mathbf{E}(\theta)$ is referred to as the energy function, and the normalizing factor $Z(\theta)$ is the partition function,

$$Z(\theta) = \sum_i \exp(-E_i(\theta)) \quad (2)$$

(here we have set the “temperature” of the system to 1).

2.2 Dynamics

Most Monte-Carlo algorithms rely on two core concepts from statistical physics, the first being conservation of probability as enforced by the master equation for the evolution of a distribution $\mathbf{p}^{(t)}$ with time [17]:

$$\dot{p}_i^{(t)} = \sum_{j \neq i} \Gamma_{ij}(\theta) p_j^{(t)} - \sum_{j \neq i} \Gamma_{ji}(\theta) p_i^{(t)}, \quad (3)$$

where $\dot{p}_i^{(t)}$ is the time derivative of $p_i^{(t)}$. Transition rates $\Gamma_{ij}(\theta)$, where $i \neq j$, give the rate at which probability flows from a state j into a state i . The first term of Equation (3) captures the flow of probability out of other states j into the state i , and the second captures flow out of i into other states j . The dependence on θ results from the requirement that the chosen dynamics cause $\mathbf{p}^{(t)}$ to flow to the equilibrium distribution $\mathbf{p}^{(\infty)}(\theta)$. For readability, explicit dependence on θ will be dropped except where necessary. If we choose to set the diagonal elements of $\mathbf{\Gamma}$ to obey $\Gamma_{ii} = -\sum_{j \neq i} \Gamma_{ji}$, then we can write the dynamics as

$$\dot{\mathbf{p}}^{(t)} = \mathbf{\Gamma} \mathbf{p}^{(t)} \quad (4)$$

(see Figure 2). The unique solution for $\mathbf{p}^{(t)}$ is²

$$\mathbf{p}^{(t)} = \exp(\mathbf{\Gamma} t) \mathbf{p}^{(0)}. \quad (5)$$

² The form chosen for $\mathbf{\Gamma}$ in Equation (4), coupled with the satisfaction of detailed balance and ergodicity introduced in section 2.3, guarantees that there is a unique eigenvector $\mathbf{p}^{(\infty)}$ of $\mathbf{\Gamma}$ with eigenvalue zero, and that all other eigenvalues of $\mathbf{\Gamma}$ have negative real parts.

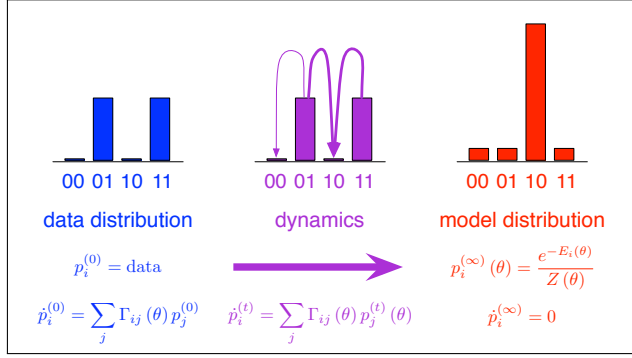


Figure 2: Dynamics of minimum probability flow learning. Model dynamics represented by the probability flow matrix Γ (*middle*) determine how probability flows from the empirical histogram of the sample data points (*left*) to the equilibrium distribution of the model (*right*) after a sufficiently long time. In this example there are only four possible states for the system, which consists of a pair of binary variables, and the particular model parameters favor state 10 whereas the data falls on other states.

2.3 Detailed Balance

The second core concept is detailed balance,

$$\Gamma_{ji} p_i^{(\infty)}(\theta) = \Gamma_{ij} p_j^{(\infty)}(\theta), \quad (6)$$

which states that at equilibrium the probability flow from state i into state j equals the probability flow from j into i . When satisfied, detailed balance guarantees that the distribution $\mathbf{p}^{(\infty)}(\theta)$ is a fixed point of the dynamics. Sampling in most Monte Carlo methods is performed by choosing Γ consistent with Equation 6 (and the added requirement of ergodicity), then stochastically running the dynamics of Equation 3. Note that there is no need to restrict the dynamics defined by Γ to those of any real physical process, such as diffusion.

Equation 6 can be written in terms of the model's energy function $\mathbf{E}(\theta)$ by substituting in Equation 1 for $\mathbf{p}^{(\infty)}(\theta)$:

$$\Gamma_{ji} \exp(-E_i(\theta)) = \Gamma_{ij} \exp(-E_j(\theta)). \quad (7)$$

Γ is underconstrained by the above equation. Motivated by symmetry and aesthetics, we choose as the form for the non-diagonal entries in Γ

$$\Gamma_{ij} = g_{ij} \exp\left[\frac{1}{2}(E_j(\theta) - E_i(\theta))\right] \quad (i \neq j), \quad (8)$$

where the connectivity function

$$g_{ij} = g_{ji} = \begin{cases} 0 & \text{unconnected states} \\ 1 & \text{connected states} \end{cases} \quad (i \neq j) \quad (9)$$

determines which states are allowed to directly exchange probability with each other³. g_{ij} can be set such that Γ is *extremely* sparse (see Section 2.5). Theoretically, to guarantee convergence to the model distribution, the non-zero elements of Γ must be chosen such that, given sufficient time, probability can flow between any pair of states.

2.4 Objective Function

Maximum likelihood parameter estimation involves maximizing the likelihood of some observations \mathcal{D} under a model, or equivalently minimizing the KL divergence between the data distribution $\mathbf{p}^{(0)}$

³The non-zero Γ may also be sampled from a proposal distribution rather than set via a deterministic scheme, in which case g_{ij} takes on the role of proposal distribution, and a factor of $\left(\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{2}}$ enters into Γ_{ij} .

and model distribution $\mathbf{p}^{(\infty)}$,

$$\hat{\theta}_{\text{ML}} = \arg \min_{\theta} D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(\infty)}(\theta) \right) \quad (10)$$

Rather than running the dynamics for infinite time, we propose to minimize the KL divergence after running the dynamics for an infinitesimal time ϵ ,

$$\hat{\theta}_{\text{MPF}} = \arg \min_{\theta} K(\theta) \quad (11)$$

$$K(\theta) = D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(\epsilon)}(\theta) \right). \quad (12)$$

For small ϵ , $D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(\epsilon)}(\theta) \right)$ can be approximated by a first order Taylor expansion,

$$K(\theta) \approx D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(t)}(\theta) \right) \Big|_{t=0} + \epsilon \frac{\partial D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(t)}(\theta) \right)}{\partial t} \Big|_{t=0}. \quad (13)$$

Further algebra (see Appendix A) reduces $K(\theta)$ to a measure of the flow of probability, at time $t = 0$ under the dynamics, out of data states \mathcal{D} into non-data states,

$$K(\theta) = \frac{\epsilon}{|\mathcal{D}|} \sum_{i \notin \mathcal{D}} \sum_{j \in \mathcal{D}} \Gamma_{ij} = \frac{\epsilon}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} g_{ij} \exp \left[\frac{1}{2} (E_j(\theta) - E_i(\theta)) \right] \quad (14)$$

with gradient⁴

$$\frac{\partial K(\theta)}{\partial \theta} = \frac{\epsilon}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} \left[\frac{\partial E_j(\theta)}{\partial \theta} - \frac{\partial E_i(\theta)}{\partial \theta} \right] g_{ij} \exp \left[\frac{1}{2} (E_j(\theta) - E_i(\theta)) \right], \quad (16)$$

where $|\mathcal{D}|$ is the number of observed data points. Note that Equations (14) and (16) do not depend on the partition function $Z(\theta)$ or its derivatives.

$K(\theta)$ is uniquely zero when $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(\infty)}(\theta)$ are exactly equal (although in general $K(\theta)$ provides a lower bound rather than an upper bound on $D_{KL} \left(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(\infty)}(\theta) \right)$). In addition, $K(\theta)$ is convex for all models $\mathbf{p}^{(\infty)}(\theta)$ in the exponential family - that is, models whose energy functions $\mathbf{E}(\theta)$ are linear in their parameters θ [13] (see Appendix B).

2.5 Tractability

The vector $\mathbf{p}^{(0)}$ is typically huge, as is $\mathbf{\Gamma}$ (e.g., 2^N and $2^N \times 2^N$, respectively, for an N -bit binary system). Naïvely, this would seem to prohibit evaluation and minimization of the objective function. Fortunately, all the elements in $\mathbf{p}^{(0)}$ not corresponding to observations are zero. This allows us to ignore all those Γ_{ij} for which no data point exists at state j . Additionally, there is a great deal of flexibility as far as which elements of \mathbf{g} , and thus $\mathbf{\Gamma}$, can be set to zero. By populating $\mathbf{\Gamma}$ so as to connect each state to a small fixed number of additional states, the cost of the algorithm in both memory and time is $\mathcal{O}(|\mathcal{D}|)$, which does not depend on the number of system states, only on the number of observed data points.

2.6 Continuous Systems

Although we have motivated this technique using systems with a large, but finite, number of states, it generalizes in a straightforward manner to continuous systems. The flow matrix $\mathbf{\Gamma}$ and distribution

⁴The contrastive divergence update rule can be written in the form

$$\Delta \theta_{CD} \propto - \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} \left[\frac{\partial E_j(\theta)}{\partial \theta} - \frac{\partial E_i(\theta)}{\partial \theta} \right] [\text{probability of MCMC step from } j \rightarrow i] \quad (15)$$

with obvious similarities to the MPF learning gradient. Thus steepest gradient descent under MPF resembles CD updates, but with the sampling/rejection step replaced by a weighting factor. Note that this difference in form provides MPF with an objective function, and guarantees a unique global minimum when model and data distributions agree.

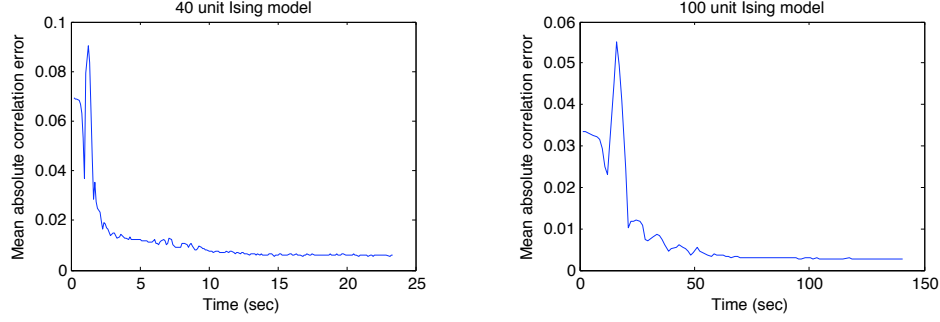


Figure 3: A demonstration of rapid fitting of the Ising model by minimum probability flow learning. The mean absolute error in the learned model’s correlation matrix is shown as a functions of learning time for 40 and 100 unit fully connected Ising models. Convergence is reached in about 15 seconds for 20,000 samples from the 40 unit model (*left*) and in about 1 minute for 100,000 samples from the 100 unit model (*right*). Details of the 100 unit model can be seen in Figure 4.

vectors $\mathbf{p}^{(t)}$ transition from being very large to being infinite in size. Γ can still be chosen to connect each state to a small, finite, number of additional states however, and only outgoing probability flow from states with data contributes to the objective function, so the cost of learning remains largely unchanged.

In addition, for a particular pattern of connectivity in Γ this objective function, like Movellan’s [15], reduces to score matching [7]. Taking the limit of connections between all states within a small distance ϵ_x of each other, and then Taylor expanding in ϵ_x , one can show that, up to an overall constant and scaling factor

$$K = K_{\text{SM}} = \sum_{i \in \mathcal{D}} \left[\frac{1}{2} \nabla E(x_i) \cdot \nabla E(x_i) - \nabla^2 E(x_i) \right]. \quad (17)$$

This reproduces the link discovered by Movellan [15] between diffusion dynamics over continuous spaces and score matching.

3 Experimental Results

Matlab code implementing minimum probability flow learning for each of the following cases is available upon request. A public toolkit is under construction.

All minimization was performed using Mark Schmidt’s remarkably effective minFunc [18].

3.1 Ising model

The Ising model has a long and storied history in physics [3] and machine learning [1] and it has recently been found to be a surprisingly useful model for networks of neurons in the retina [19, 21]. The ability to fit Ising models to the activity of large groups of simultaneously recorded neurons is of current interest given the increasing number of these types of data sets from the retina, cortex and other brain structures.

We fit an Ising model (fully visible Boltzmann machine) of the form

$$p^{(\infty)}(\mathbf{x}; \mathbf{J}) = \frac{1}{Z(\mathbf{J})} \exp \left[- \sum_{i,j} J_{ij} x_i x_j \right] \quad (18)$$

to a set of N d -element iid data samples $\{x^{(i)} | i = 1 \dots N\}$ generated via Gibbs sampling from an Ising model as described below, where each of the d elements of \mathbf{x} is either 0 or 1. Because each $x_i \in \{0, 1\}$, $x_i^2 = x_i$, we can write the energy function as

$$E(\mathbf{x}; \mathbf{J}) = \sum_{i,j \neq i} J_{ij} x_i x_j + \sum_i J_{ii} x_i. \quad (19)$$

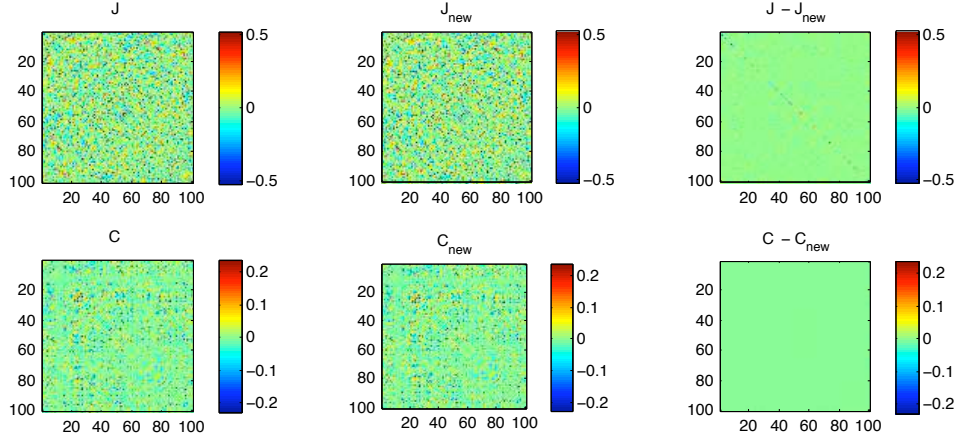


Figure 4: An example 100 unit Ising model fit using minimum probability flow learning. (*left*) Randomly chosen Gaussian coupling matrix J (top) with variance 0.04 and associated correlation matrix C (bottom) for a 100 unit, fully-connected Ising model. The diagonal has been removed from the correlation matrix C for increased visibility. (*center*) The recovered coupling and correlation matrices after minimum probability flow learning on 100,000 samples from the model in the left panels. (*right*) The error in recovery of the coupling and correlation matrices.

The probability flow matrix Γ has $2^N \times 2^N$ elements, but for learning we populate it extremely sparsely, setting

$$g_{ij} = g_{ji} = \begin{cases} 1 & \text{states } i \text{ and } j \text{ differ by single bit flip} \\ 0 & \text{otherwise} \end{cases}. \quad (20)$$

Figure 3 shows the average error in predicted correlations as a function of learning time for 20,000 samples from a 40 unit, fully connected Ising model. The J_{ij} used were graciously provided by Broderick and coauthors, and were identical to those used for synthetic data generation in the 2008 paper “Faster solutions of the inverse pairwise Ising problem” [2]. Training was performed on 20,000 samples so as to match the number of samples used in section III.A. of Broderick et al. Note that given sufficient samples, the minimum probability flow algorithm would converge exactly to the right answer, as learning in the Ising model is convex (see Appendix B), and has its global minimum at the true solution. On an 8 core 2.33 GHz Intel Xeon, the learning converges in about 15 seconds. Broderick et al. perform a similar learning task on a 100-CPU grid computing cluster, with a convergence time of approximately 200 seconds.

Similar learning was performed for 100,000 samples from a 100 unit, fully connected Ising model. A coupling matrix was chosen with elements randomly drawn from a Gaussian with mean 0 and variance 0.04. Using the minimum probability flow learning technique, learning took approximately 1 minute, compared to roughly 12 hours for a 100 unit (nearest neighbor coupling only) model of retinal data [20] (personal communication, J. Shlens). Figure 4 demonstrates the recovery of the coupling and correlation matrices for our fully connected Ising model, while Figure 3 shows the time course for learning.

3.2 Deep Belief Network

As a demonstration of learning on a more complex discrete valued model, we trained a 4 layer deep belief network (DBN) [6] on MNIST handwritten digits. A DBN consists of stacked restricted Boltzmann machines (RBMs), such that the hidden layer of one RBM forms the visible layer of the



Figure 5: A deep belief network trained using minimum probability flow learning and contrastive divergence. (*left*) A four layer deep belief network was trained on the MNIST postal hand written digits dataset. (*center*) Confabulations after training via minimum probability flow learning. A reasonable probabilistic model for handwritten digits has been learned. (*right*) Confabulations after training via single step CD. Note the uneven distribution of digit occurrences.

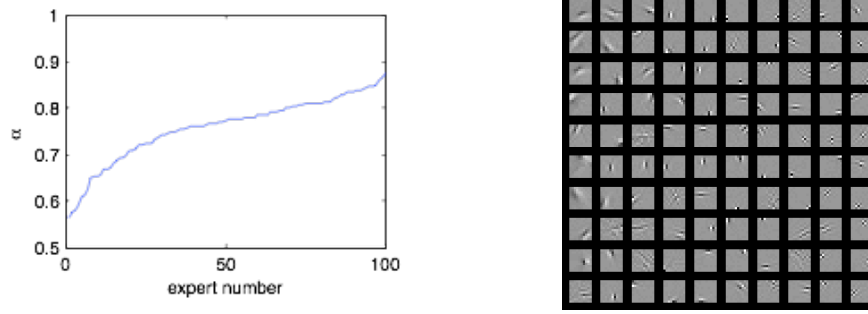


Figure 6: A continuous state space model fit using minimum probability flow learning. The sparsity parameter α (*left*) and receptive fields \mathbf{J} (*right*) are shown for a product of Student-t model trained on natural images. Both plots are ordered by sparsity.

next. Each RBM has the form:

$$p^{(\infty)}(\mathbf{x}_{\text{vis}}, \mathbf{x}_{\text{hid}}; \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp \left[- \sum_{i,j} W_{ij} x_{\text{vis},i} x_{\text{hid},j} \right], \quad (21)$$

$$p^{(\infty)}(\mathbf{x}_{\text{vis}}; \mathbf{W}) = \frac{1}{Z(\mathbf{W})} \exp \left[\sum_j \log \left(1 + \exp \left[- \sum_i W_{ij} x_{\text{vis},i} \right] \right) \right]. \quad (22)$$

Note that sampling-free application of the minimum probability flow algorithm requires analytically marginalizing over the hidden units. RBMs were trained in sequence, starting at the bottom layer, on 10,000 samples from the MNIST postal hand written digits data set. As in the Ising case, the probability flow matrix $\mathbf{\Gamma}$ was populated so as to connect every state to all states which differed by only a single bit flip (see Equation 20). Training was performed by both minimum probability flow and single step CD to allow a simple comparison of the two techniques (note that CD turns into full ML learning as the number of steps is increased, and that the quality of the CD answer can thus be improved at the cost of computational time by using many-step CD).

Confabulations were performed by Gibbs sampling from the top layer RBM, then propagating each sample back down to the pixel layer by way of the conditional distribution $p^{(\infty)}(\mathbf{x}_{\text{vis}} | \mathbf{x}_{\text{hid}}; \mathbf{W}^k)$ for each of the intermediary RBMs, where k indexes the layer in the stack. As shown in Figure 5, minimum probability flow learned a good model of handwritten digits.

3.3 Product of Student-t Model

As a demonstration of parameter estimation in continuous state space, non-analytically normalizable, probabilistic models, we trained a product of Student-t model [25] with 100 receptive fields \mathbf{J}

on 100,000 10×10 natural image patches. The product of Student-t model has the form

$$p^{(\infty)}(\mathbf{x}; \mathbf{J}, \alpha) \propto e^{-\sum_i \alpha_i \log[1 + (\mathbf{J}_i \mathbf{x})^2]}, \quad (23)$$

where the α parameter controls the sparsity of the distribution. g_{ij} was chosen such that each state \mathbf{x} was connected to 2 states $\tilde{\mathbf{x}}$ chosen by adding Gaussian noise to \mathbf{x} and rescaling to maintain patch variance — that is $\tilde{\mathbf{x}} = (\mathbf{x} + \tilde{\mathbf{n}}) \frac{\|\mathbf{x}\|_2}{\|\mathbf{x} + \tilde{\mathbf{n}}\|_2}$ for Gaussian noise vector $\tilde{\mathbf{n}} \sim N(0, 0.1)$. Figure 6 shows that this model learns oriented edges, as expected.

4 Summary

We have presented a novel framework for efficient learning in the context of any parametric model. This method was inspired by the minimum velocity approach developed by Movellan, and it reduces to that technique as well as to score matching and some forms of contrastive divergence under suitable choices for the dynamics and state space. By decoupling the dynamics from any specific physical process, such as diffusion, and focusing on the initial flow of probability from the data to a subset of other states chosen in part for their utility and convenience, we have arrived at a framework that is not only more general than previous approaches, but also potentially much more powerful. We expect that this framework will render some previously intractable models more amenable to analysis.

Acknowledgments

We would like to thank Javier Movellan for sharing a work in progress; Tamara Broderick, Miroslav Dudík, Gašper Tkačik, Robert E. Schapire and William Bialek for use of their Ising model coupling parameters; Jonathon Shlens for useful discussion and ground truth for his Ising model convergence times; Bruno Olshausen, Anthony Bell, Christopher Hillar, Charles Cadieu, Kilian Koepsell and the rest of the Redwood Center for many useful discussions and for comments on earlier versions of the manuscript; Ashvin Vishwanath for useful discussion; and the Canadian Institute for Advanced Research - Neural Computation and Perception Program for their financial support (JSD).

APPENDICES

A Taylor Expansion of KL Divergence

$$K(\theta) \approx D_{KL}(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(t)}(\theta)) \Big|_{t=0} + \epsilon \frac{\partial D_{KL}(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(t)}(\theta))}{\partial t} \Big|_{t=0} \quad (\text{A-1})$$

$$= 0 + \epsilon \frac{\partial D_{KL}(\mathbf{p}^{(0)} \parallel \mathbf{p}^{(t)}(\theta))}{\partial t} \Big|_{t=0} \quad (\text{A-2})$$

$$= \epsilon \frac{\partial}{\partial t} \left(\sum_{i \in \mathcal{D}} p_i^{(0)} \log \frac{p_i^{(0)}}{p_i^{(t)}} \right) \Big|_0 \quad (\text{A-3})$$

$$= -\epsilon \sum_{i \in \mathcal{D}} \frac{p_i^{(0)}}{p_i^{(0)}} \frac{\partial p_i^{(t)}}{\partial t} \Big|_0 \quad (\text{A-4})$$

$$= -\epsilon \sum_{i \in \mathcal{D}} \frac{\partial p_i^{(t)}}{\partial t} \Big|_0 \quad (\text{A-5})$$

$$= -\epsilon \left(\frac{\partial}{\partial t} \sum_{i \in \mathcal{D}} p_i^{(t)} \right) \Big|_0 \quad (\text{A-6})$$

$$= -\epsilon \frac{\partial}{\partial t} \left(1 - \sum_{i \notin \mathcal{D}} p_i^{(t)} \right) \Big|_0 \quad (\text{A-7})$$

$$= \epsilon \sum_{i \notin \mathcal{D}} \frac{\partial p_i^{(t)}}{\partial t} \Big|_0 \quad (\text{A-8})$$

$$= \epsilon \sum_{i \notin \mathcal{D}} \sum_{j \in \mathcal{D}} \Gamma_{ij} p_j^{(0)} \quad (\text{A-9})$$

$$= \frac{\epsilon}{|\mathcal{D}|} \sum_{i \notin \mathcal{D}} \sum_{j \in \mathcal{D}} \Gamma_{ij}, \quad (\text{A-10})$$

where we used the fact that $\sum_{i \in \mathcal{D}} p_i^{(t)} + \sum_{i \notin \mathcal{D}} p_i^{(t)} = 1$. This implies that the rate of growth of the KL divergence at time $t = 0$ equals the total initial flow of probability from states with data into states without.

B Convexity

As observed by Macke and Gerwinn [13], the MPF objective function is convex for models in the exponential family.

We wish to minimize

$$K = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}^c} \Gamma_{ji} p_i^{(0)}. \quad (\text{B-1})$$

K has derivative

$$\frac{\partial K}{\partial \theta_m} = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}^c} \left(\frac{\partial \Gamma_{ij}}{\partial \theta_m} \right) p_i^{(0)} \quad (\text{B-2})$$

$$= \frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}^c} \Gamma_{ij} \left(\frac{\partial E_j}{\partial \theta_m} - \frac{\partial E_i}{\partial \theta_m} \right) p_i^{(0)}, \quad (\text{B-3})$$

and Hessian

$$\begin{aligned} \frac{\partial^2 K}{\partial \theta_m \partial \theta_n} &= \frac{1}{4} \sum_{i \in D} \sum_{j \in D^c} \Gamma_{ij} \left(\frac{\partial E_j}{\partial \theta_m} - \frac{\partial E_i}{\partial \theta_m} \right) \left(\frac{\partial E_j}{\partial \theta_n} - \frac{\partial E_i}{\partial \theta_n} \right) p_i^{(0)} \\ &\quad + \frac{1}{2} \sum_{i \in D} \sum_{j \in D^c} \Gamma_{ij} \left(\frac{\partial^2 E_j}{\partial \theta_m \partial \theta_n} - \frac{\partial^2 E_i}{\partial \theta_m \partial \theta_n} \right) p_i^{(0)}. \end{aligned} \quad (\text{B-4})$$

The first term is a weighted sum of outer products, with non-negative weights $\frac{1}{4} \Gamma_{ij} p_i^{(0)}$, and is thus positive semidefinite. The second term is 0 for models in the exponential family (those with energy functions linear in their parameters).

Parameter estimation for models in the exponential family is therefore convex using minimum probability flow learning.

References

- [1] D H Ackley, G E Hinton, and T J Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(2):147–169, 1985.
- [2] T Broderick, M Dudík, G Tkačik, R Schapire, and W Bialek. Faster solutions of the inverse pairwise Ising problem. *E-print arXiv*, Jan 2007.
- [3] S G Brush. History of the Lenz-Ising model. *Reviews of Modern Physics*, 39(4):883–893, Oct 1967.
- [4] M A Carreira-Perpiñán and G E Hinton. On contrastive divergence (CD) learning. *Technical report, Dept. of Computer Science, University of Toronto*, 2004.
- [5] S Haykin. *Neural networks and learning machines; 3rd edition*. Prentice Hall, 2008.
- [6] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, Jul 2006.
- [7] A Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
- [8] A Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, Jan 2007.
- [9] T Jaakkola and M Jordan. A variational approach to Bayesian logistic regression models and their extensions. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Jan 1997.
- [10] H Kappen and F Rodríguez. Mean field approach to learning in Boltzmann machines. *Pattern Recognition Letters*, Jan 1997.
- [11] S Lyu. Interpretation and generalization of score matching. *The proceedings of the 25th conference on uncertainty in artificial intelligence (UAI*90)*, 2009.
- [12] D MacKay. Failures of the one-step learning algorithm. <http://www.inference.phy.cam.ac.uk/mackay/gbm.pdf>, Jan 2001.
- [13] J Macke and S Gerwinn. Personal communication. 2009.
- [14] J R Movellan. Contrastive divergence in Gaussian diffusions. *Neural Computation*, 20(9):2238–2252, 2008.
- [15] J R Movellan. A minimum velocity approach to learning. *unpublished draft*, Jan 2008.
- [16] J R Movellan and J L McClelland. Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, 17:463–496, 1993.
- [17] R Pathria. *Statistical Mechanics*. Butterworth Heinemann, Jan 1972.
- [18] M Schmidt. minfunc. <http://www.cs.ubc.ca/~schmidt/Software/minFunc.html>, 2005.
- [19] E Schneidman, M J Berry 2nd, R Segev, and W Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–12, 2006.
- [20] J Shlens, G D Field, J L Gauthier, M Greschner, A Sher, A M Litke, and E J Chichilnisky. The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience*, 29(15):5022–5031, Apr 2009.
- [21] J Shlens, G D Field, J L Gauthier, M I Grivich, D Petrusca, A Sher, A M Litke, and E J Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.*, 26(32):8254–66, 2006.
- [22] J Sohl-Dickstein and B Olshausen. A spatial derivation of score matching. *Redwood Center Technical Report*, 2009.

- [23] T Tanaka. Mean-field theory of Boltzmann machine learning. *Physical Review Letters E*, Jan 1998.
- [24] M Welling and G Hinton. A new learning algorithm for mean field Boltzmann machines. *Lecture Notes in Computer Science*, Jan 2002.
- [25] M Welling, G Hinton, and S Osindero. Learning sparse topographic representations with products of student-t distributions. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, Jan 2003.
- [26] A Yuille. The convergence of contrastive divergences. *Department of Statistics, UCLA. Department of Statistics Papers.*, 2005.